HumRRO Report
FR-02-47
**September, 2002**

# Third-Party Checking of 2002 Scaling and Equating for the Kentucky Core Content Test

**Arthur A. Thacker**
**Laura A. Ford**
**R. Gene Hoffman**

**Prepared for:**

**The Kentucky Department of Education**
**500 Mero Street**
**Frankfort, KY 40601**

**Contract Number M-00003669**

*Human Resources Research Organization*
66 Canal Center Plaza, Suite 400 • Alexandria, VA 22314-1591

# Third-Party Checking of 2002 Scaling and Equating for the Kentucky Core Content Test

## Table of Contents

## Summary

CTB and HumRRO independently calculated the scaled and equated raw-score-to-scale-score tables for the 2002 Kentucky Core Content Test. From those tables, cut points were identified that can be used for (1) assigning student performance classifications and (2) converted to school accountability indexes. Decisions regarding the handling of problem test items were discussed between CTB and HumRRO and in all cases both groups reached consensus. All results calculated by HumRRO were identical to those calculated by CTB. Given that our scaling and equating results were identical with those of CTB, we are assured that CTB did not commit processing errors.

# Third-Party Checking of 2002 Scaling and Equating for the Kentucky Core Content Test

## Introduction

Every year, the Kentucky Core Content Test (KCCT) [1] is scaled and equated by Item Response Theory (IRT) using a calibration sample of students in designated grades (4, 5, 7, 8, 10, and 11). Scaling involves the estimation of item parameters for the current year's test. These item parameters are then linearly transformed to a 325-800 point scale (mean = 500, standard deviation = 50) and equated with the previous year's scale. The results of scaling and equating are then used to construct raw-score-to-scale-score tables for every KCCT test form. Cut points are also identified so that students' raw scores can be translated to performance categories: Novice, Apprentice, Proficient, and Distinguished.

This process allows Kentucky to maintain consistency in its student performance levels from year to year. During the transition from the Kentucky Instructional Results Information System (KIRIS) test in 1997 to the KCCT in 1998 a system was devised (see Hoffman, Thacker, & Ford, 2001; Hoffman & Thacker, 2000; Hoffman & Thacker, 1999) to link KCCT item parameters and KIRIS item parameters on the same scale. The result is that year-to-year KCCT student performance levels from 1998-2002 can be compared to KIRIS performance levels as far back as 1992. Scaling and equating are done for the following grade/subject combinations:

Grade 4 - Reading, Science
Grade 5 - Math, Social Studies, Arts & Humanities, Practical Living/Vocational Studies
Grade 7 - Reading, Science
Grade 8 - Math, Social Studies, Arts & Humanities, Practical Living/Vocational Studies
Grade 10 - Reading, Practical Living/Vocational Studies
Grade 11 - Math, Science, Social Studies, Arts & Humanities

As a quality control step, personnel at CTB and the Human Resources Research Organization (HumRRO) conduct scaling and equating analyses simultaneously and independently. Researchers at both companies compare results at several steps throughout the process. If a result between CTB and HumRRO is not identical, then procedures are reviewed until the issue is resolved and both staffs get the same outcome. This way, the complex sampling, item parameter estimation analyses, Stocking-Lord equating, raw-score-to-scale-score transformations, and cut points identifications are checked and verified by two, autonomous agencies.

The procedures used by HumRRO are outlined in detail below.

[1] The test in use before 1998 was the Kentucky Instructional Results Information System (KIRIS) test.

## Scaling and Equating Procedures

Item response data for all of the 2002 test forms were scaled using CTB's PARDUX program. Based on IRT, PARDUX uses a three-parameter logistic model for multiple-choice items and a two-parameter model for open-response items to estimate item parameters. Item parameters from both these models are eventually transformed to a single scale.

The equating process involves the application of the Stocking-Lord procedure to two different sets of anchor[2] item parameters: anchor item parameters from 2001 and anchor item parameters from 2002. These two sets of parameters are on different metrics. The 2002 parameters are on a theta metric (-1 to +1 scale) and the 2001 item parameters are on the "Kentucky metric" (325 to 800 scale). Stocking-Lord produces transformation constants (M1 and M2) that are used to linearly transform the 2002 metric onto the 2001 metric. This transforms all the 2002 item parameters onto the 325 – 800 scale, which traces back to the original 1992 scale.

The final step in the process is to use CTB's FLUX program to create raw-score-to-scale-score conversion tables and identify the cut points for the performance categories.

## Scope of Third-Party Checking

In addition to doing a parallel analysis with CTB this year, HumRRO also conducted in-house, parallel analysis to accomplish scaling and linking for the 2002 data. The Processing Steps listed below, while adequate, are being improved each year to ensure greater accuracy, standardization, and efficiency. This year, HumRRO developed and tested a new SAS program that automates much of the first 3 steps. The two main improvements provided by this program are: (1) it automatically generates anchor files for all grades from the previous year's PARDUX*.par files, and (2) it verifies the calibration sample by filtering students according to the eligibility criteria and comparing it to the calibration sample produced by CTB. One HumRRO researcher followed the procedure outlined below and two other researchers tried out the new SAS program.

## Processing Steps

HumRRO took the following steps for each grade/subject tested:

1. Created anchor files (PARDUX *.anc) of multiple-choice test items that appeared on the anchor form. These anchor items were used to equate the 2002 test to the 2001 scale. The 2002 anchor files were created using 2001 parameter files for the matching forms.

2. Created working files (PARDUX *.RWO) from the calibration sample for the 2002 Kentucky Core Content Test. These files include both open-response and multiple-choice data.

[2] Anchor items were designated on one form for each grade/subject on the 2001 KCCTs. The same anchor form was readministered in 2002 with all items intact and occurring in the same sequence as in 2001.

3. Prepared control files (PARDUX *.ctl) which contain the constraints used for item parameter estimation, student proficiency estimation, maximum number of items, etc. The SAS program used to create *.rwo files included a routine to print out a control file.

4. Estimated parameters for Kentucky Core Content Test items using PARDUX.

5. Performed Stocking-Lord transformation using PARDUX. The results of this transformation include a slope and intercept constant for equating the 2002 Kentucky Core Content Test back to 2001.

6. Confirmed that the equating constants (M1 and M2) from Step 5 match those derived by CTB.

7. Created parameter files (FLUX *.par) for each test form for use in preparation of raw-score-to-scale-score tables. A special SAS program was written for this purpose.

8. Created files (FLUX *.hlk) containing the scale limits (325 and 800) and constants from the Stocking-Lord transformation. This was a simple word processing task.

9. Created raw-score-to-scale-score transformation tables for each form using FLUX.

10. Confirmed that the raw-score-to-scale-score transformation tables from Step 9 match those derived by CTB and verified cut points used to separate student performance into Novice (Non-performing, Middle, High)/Apprentice (Low, Middle, High)/Proficient/Distinguished categories.

## Results

After performing periodic checks with CTB as individual tests were scaled and equated, HumRRO and CTB reached exact agreement on the equating constants for all grade/subjects. Table 1 summarizes the results of this study. Grade and subject are identified for each test in the first two columns, respectively. The stage at which convergence occurred (if at all) is recorded in the third column. The fourth column identifies problem items and references the solutions that were reached by CTB and verified by HumRRO. The next four columns contain the M1 and M2 (slope and intercept) constants obtained from the Stocking-Lord transformation. CTB computed the first set of constants and HumRRO the second. The ninth column contains the difference between CTB's and HumRRO's M1 constants (i.e., $M1_{CTB}$-$M1_{HumRRO}$). The tenth column records the same information for M2 constants (i.e., $M2_{CTB}$-$M2_{HumRRO}$).

The last two columns in Table 1 list whether there was exact agreement between CTB and HumRRO on (1) the raw-score-to-scale-score tables and (2) the cut points. Cut points from these tables are used to assign students to performance categories that, in turn, are used in the computation of each school's accountability index. CTB and HumRRO were in exact agreement for all raw-score-to-scale-score tables for every grade/subject.

Explanations of convergence issues and individual item issues are footnoted in Table 1. The footnotes explain the specific problems and their solutions. It should be noted that all problem

items were dealt with during the parameter estimation phase of the scaling and equating process. No item for which parameters were estimated was eliminated from the Stocking-Lord procedure. The same column indicates whether or not convergence was reached during parameter estimation. If convergence was not reached after 50 iterations by the PARDUX program, the solution at stage 50 was accepted by mutual agreement.

Table 1

Comparison of HumRRO and CTB Scaling and Linking Results

| | | | | CTB | | HUMRRO | | CTB-HUMRRO Differences | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Grade* | *Subject* | *Convergence* | *Problems* | *M1* | *M2* | *M1* | *M2* | *M1* | *M2* | *Tables Agree* | *Cut points check* |
| 4 | RD | Stage 12 | None | 31.10756 | 548.90155 | 31.10756 | 548.90155 | 0.00000 | 0.00000 | Yes | Yes |
| | SC | Stage 14 | None | 24.68294 | 548.06689 | 24.68294 | 548.06689 | 0.00000 | 0.00000 | Yes | Yes |
| 5 | A&H | Stage 17 | None | 44.69901 | 517.86450 | 44.69901 | 517.86450 | 0.00000 | 0.00000 | Yes | Yes |
| | MA | Stage 17 | None | 33.98995 | 562.16522 | 33.98995 | 562.16522 | 0.00000 | 0.00000 | Yes | Yes |
| | PL | Stage 14 | None | 44.19806 | 507.33032 | 44.19806 | 507.33032 | 0.00000 | 0.00000 | Yes | Yes |
| | SS | Stage 15 | None | 30.95921 | 541.33759 | 30.95921 | 541.33759 | 0.00000 | 0.00000 | Yes | Yes |
| 7 | RD | No[1] | Convergence, Item 81[3] | 29.01876 | 514.81335 | 29.01876 | 514.81335 | 0.00000 | 0.00000 | Yes | Yes |
| | SC | Stage 21 | None | 26.41763 | 505.17697 | 26.41763 | 505.17697 | 0.00000 | 0.00000 | Yes | |
| 8 | A&H | Stage 35 | Items 50 & 120[4] | 46.92245 | 512.30212 | 46.92245 | 512.30212 | 0.00000 | 0.00000 | Yes | Yes |
| | MA | Stage 15 | None | 31.69082 | 533.87628 | 31.69082 | 533.87628 | 0.00000 | 0.00000 | Yes | Yes |
| | PL | Stage 16 | Item 99[5] | 40.15432 | 502.33768 | 40.15432 | 502.33768 | 0.00000 | 0.00000 | Yes | Yes |
| | SS | Stage 21 | None | 39.96317 | 514.78442 | 39.96317 | 514.78442 | 0.00000 | 0.00000 | Yes | Yes |
| 10 | PL | Stage 19 | None | 46.76192 | 503.85443 | 46.76192 | 503.85443 | 0.00000 | 0.00000 | Yes | Yes |
| | RD | Stage 14 | None | 52.32949 | 505.28320 | 52.32949 | 505.28320 | 0.00000 | 0.00000 | Yes | Yes |
| 11 | A&H | Stage 23 | Item 100[6] | 52.44464 | 524.14539 | 52.44464 | 524.14539 | 0.00000 | 0.00000 | Yes | Yes |
| | MA | No[2] | Convergence | 41.08334 | 536.37134 | 41.08334 | 536.37134 | 0.00000 | 0.00000 | Yes | Yes |
| | SC | Stage 20 | None | 31.49916 | 546.50195 | 31.49916 | 546.50195 | 0.00000 | 0.00000 | Yes | Yes |
| | SS | Stage 15 | Items 96, 98, & 127[7] | 49.55783 | 548.98932 | 49.55783 | 548.98932 | 0.00000 | 0.00000 | Yes | Yes |

[1] Convergence was not reached for RD07. The solution at Stage 50 was used operationally.
[2] Convergence was not reached for MA11. The solution at Stage 50 was used operationally.
[3] Item 81 had some extreme parameters, but removing it did not significantly change M1 and M2, so the item remained.
[4] Items 50 and 120 in AH08 both required an M-step for parameter estimation.
[5] Item 99 in PL08 required an M-step for parameter estimation.
[6] Item 100 in AH11 required an M-step for parameter estimation.
[7] Items 96, 98, and 127 in SS11 all required an M-step for parameter estimation.

HumRRO also verified the cut points on the raw-score-to-scale-score tables. Cut points were assigned by rule. HumRRO verified cut points between Novice and Apprentice, between Apprentice and Proficient, and between Proficient and Distinguished performance categories. HumRRO also verified cut points for Low, Medium, and High subcategories within the Novice and Apprentice categories.

## Documentation

To document the steps involved in scaling and linking the 2002 Kentucky Core Content Test HumRRO saved all electronic files used in data preparation, including SAS programs, SAS logs, and SAS output lists and all files produced during PARDUX scaling and FLUX transformations. These files have been submitted to Kentucky Department of Education (KDE). Appendices from the Hoffman and Thacker (1999) report contain hardcopy examples of important files that were submitted.

All electronic files submitted to KDE are named according to the following code (where S = subject, G = grade level).

A. PARDUX Control File (SSGG02.CTL). This file contains the number of items, the maximum number of stages for PARDUX, the convergence criterion, parameter estimation limits, maximum and minimum values for proficiency estimates (theta). It also contains information allowing the program to distinguish between open-response and multiple-choice items, the number of score levels for open-response data, and which items to include in parameter estimation.

B. PARDUX Data File (SSGG02.RWO). This file contains the student score data. It is coded such that a 1 indicates a correct answer for a multiple-choice question and actual score levels (0-4) are recorded for student responses to open-response questions. To facilitate communication, HumRRO adhered to CTB's item order in constructing these data files.

C. PARDUX Anchor File (SSGG02.ANC). This file contains common-scaling item parameters from the 2001 KCCT (the identical items appeared on the 2002 KCCT). Only multiple-choice items are used in *.ANC files.

D. SAS Programs configured as SSGGrwcd.sas. This program produces the anchor files (*.ANC), PARDUX control files (*.CTL), and student score files (*.RWO). The SAS log and list files generated by these programs are also included electronically.

E. SAS Programs configured as SSGGmakeparfiles.sas. For each grade-subject, this program sorts the parameter data by test form, a configuration required by the FLUX program.

F. PARDUX Parameter Estimation Summary (SSGG02_SUM.TXT). This file provides a summary of the parameter estimation procedure run in PARDUX. It includes the limit data from the control file and also contains the number of stages PARDUX ran in order to reach convergence. It also contains the item numbers of items that could not be estimated and documents any items whose estimation reaches the maximum alpha parameter. This file identifies any problem items that might require additional manipulation before continuing the process.

G. PARDUX Parameter Estimation Details (SSGG02_DET.TXT). This file lists a systematic iteration of data, by item, during each stage of parameter estimation.

H. PARDUX Parameter File (SSGG02.PAR). This file contains parameter estimates for all items designated in the *.CTL file. It is used for later data manipulation.

I. PARDUX Item Summaries Files, Status (SSGG02_STAT.TXT). This file lists all items for a given test and their status after parameter estimation. Items are coded as either "estimate OK," "OK—default C," "not estimated," or "other codes." It provides a different type of record for the parameter estimation.

J. PARDUX Item Summaries Files, Distribution (SSGG02_DIST.TXT). This file contains the distribution of students who scored at each level on the open-response items. It is useful for examining the way that scoring rubrics for these items operate and for ensuring that all open-response items have the correct number of functioning score levels.

K. PARDUX Item Summaries Files, Parameters (SSGG02_PAR.TXT). This file contains the item parameters in different format from the *.PAR files. Word processing and spreadsheet programs can easily read this file.

L. PARDUX Item Summaries Files, Standard Errors (SSGG02_SE.TXT). This file contains the standard errors of estimation for each item including the errors for the various score levels on the open response items.

M. PARDUX Item Summaries Files, FitQ1 (SSGG02_Q1.TXT). This file contains fit statistics for all items.

N. PARDUX Log File (SSGG02_LOG.TXT). As each manipulation of data is completed, PARDUX maintains a log of the procedures and filenames. This log is saved in text format.

O. Stocking-Lord Plots (SSGG02_SLPLOTS.doc). For each grade/subject combination, the Stocking-Lord data transformation calculates M1 and M2 values (slope and intercept) and outputs four graphs (one each for the a, b, and c parameters, and item p-values). The M1/M2 values, a log of the Stocking-Lord procedures, and the graphs are saved in this file.

P. FLUX control file (SSGG02.HLK). This file specifies the range of the scale scores as well as the M1 and M2 transformation constants from the Stocking-Lord transformation.

Q. FLUX Parameter Files by Form (SSGG021A.PAR, SSGG021B.PAR, etc.). Each parameter file computed using PARDUX was divided to represent items from each test form. Typically, 30 items were scored from each form. The exceptions are forms from Arts and Humanities and Practical Living/Vocational Studies, which each contain only 10 scored items.

R. Raw-Score-to-Scale-Score Tables (SSGG02RStoSSTables.doc). A raw-score-to-scale-score table was produced for each form. These tables were saved in text format using FLUX.

S. Miscellaneous files and programs may also be included in the documentation. These files were constructed either during investigation of results or for future purposes. Student data records (provided by DRC) from which all 2002 data were extracted are included as well.

## Conclusion

CTB and HumRRO independently calculated the scaled/equated raw-score-to-scale-score tables for the 2002 Kentucky Core Content Test. From these tables, both identified cut points that could be used for assigning student performance classifications and later converted to school accountability indexes. No differences were found between CTB's and HumRRO's parameter estimation, Stocking-Lord transformation constants, raw-score-to-scale-score tables, or application of cut points. In addition, only slight differences were found using the "old" 1-3 Processing Steps and the new HumRRO SAS program. The differences that were found were in rounding of anchor item parameters – these rounding differences were so small that they had negligible effect on M1/M2 values and no effect on final cut points.

Given that the HumRRO and CTB scaling and linking results were identical, HumRRO is confident that CTB did not commit processing errors.

## References

Hoffman, R. G., Thacker, A. A, & Ford. L. A. (2001). *Third-party checking of 2001 scaling and linking for the Kentucky Core Content Test*. (HumRRO Report FR-01-41). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G. & Thacker, A. A. (2000). *Third-party checking of 2000 scaling and linking for the Kentucky Core Content Test*. (HumRRO Report FR-00-39). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G. & Thacker, A. A. (1999). *Third-party checking of 1999 scaling and linking for the Kentucky Core Content Test*. (HumRRO Report SP-WATSD-99-44). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G., Thacker, A. A., & McBride, J. R. (1999). *Documentation of third-party checking of 1998 pre-equating for Kentucky Core Content Test: IRT scaling of multiple choice and open response test items.* (HumRRO Report SP-WATSD-99-39). Alexandria, VA: Human Resources Research Organization.